ehp Environmental Health PERSPECTIVES
ehponline.org
Published by the National Institute of Environmental Health Sciences

Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits

Jay H. Lubin, Joanne S. Colt, David Camann, Scott Davis, James R. Cerhan, Richard K. Severson, Leslie Bernstein, and Patricia Hartge

doi:10.1289/ehp.7199 (available at http://dx.doi.org/)
Online 13 September 2004

The National Institute of Environmental Health Sciences
National Institutes of Health
U.S. Department of Health and Human Services

Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits

Jay H. Lubin[1], Joanne S. Colt[1], David Camann[2], Scott Davis[3], James R. Cerhan[4]
Richard K. Severson[5], Leslie Bernstein[6], Patricia Hartge[1]

[1] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD
[2] Southwest Research Institute, San Antonio, TX
[3] Fred Hutchinson Cancer Research Center and the University of Washington, Seattle, WA
[4] Mayo Clinic, College of Medicine, Rochester, MN
[5] Karmanos Cancer Institute and Department of Family Medicine, Wayne State University, Detroit, MI
[6] Department of Preventive Medicine, Norris Comprehensive Cancer Center, Keck School at Medicine at the University of Southern California, Los Angeles, CA

Corresponding author:

Jay Lubin, National Cancer Institute, Biostatistics Branch, 6120 Executive Boulevard, Room 8042, Rockville, MD 20852

Phone number: (301) 496-3357
FAX number: (301) 402-0081
E-mail: lubinj@mail.nih.gov.

Running title: Measurement data with detection limits

Keywords: pesticides, dust, environmental exposure, imputation, missing data, non-Hodgkin lymphoma

Acknowledgment and support:
The authors declare they have no competing financial interest. Support for this study included contracts with the National Cancer Institute: N01-PC-67010, N01-PC-67008, N02-PC-71105, N01-PC-67009, N01-PC-65064.

Abbreviations:

AM	Arithmetic mean
CI	Confidence interval
DL	Detection limit
GC/MS	Gas chromatography/mas spectrometry
GM	Geometric mean
GSD	Geometric standard deviation
LB	Lower bound
MLE	Maximum likelihood estimate
ng/g	Nano-grams per gram
NHL	Non-Hodgkin lymphoma
SE	Standard error
UB	Upper bound

Outline of section headings:

Abstract

Quantitative measurements of environmental factors greatly improve the quality of epidemiologic studies, but can pose challenges due to the presence of upper or lower detection limits or interfering compounds, which do not allow for precise measured values. We consider the regression of an environmental measurement (dependent variable) on several covariates (independent variables). Various strategies are commonly employed to impute values for interval-measured data, including assignment of one-half the detection limit to non-detected values, or of "fill-in" values randomly selected from an appropriate distribution. Based on a limited simulation study, we found that the former approach can be biased, unless the percentage of measurements below detection limits is small (5-10 percent). The fill-in approach generally results in unbiased parameter estimates, but may produce biased variance estimates and thereby distort inference when 30 percent or more of the data are below detection limits. Truncated data methods (e.g., Tobit regression) and multiple imputation offer two unbiased approaches for analyzing measurement data with detection limits. If interest resides solely on regression parameters, then Tobit regression can be used. If individualized values for measurements below detection limits are needed for additional analysis, such as relative risk regression or graphical display, then multiple imputation produces unbiased estimates and nominal confidence intervals unless the proportion of missing data is extreme. We illustrate various approaches using measurements of pesticide residues in carpet dust in control subjects from a case-control study of non-Hodgkin lymphoma.

Introduction

Epidemiologic studies often collect quantitative measurement data to improve precision and reduce bias in exposure assessment, and in the estimation of the effect of exposure on risk of disease, as measured by odds ratios (Hatch and Thomas 1993; Sim 2002). Some measurements serve as biomarkers for "dose", e.g., residual radiation in tooth enamel as a marker of exposure to ionizing radiation (Desrosiers and Schauer 2001), while other measures are more indirect, e.g., urinary cotinine level as an indicator of exposure to environmental tobacco smoke (Woodward and Al Delaimy 1999). Problems in the analysis of measurement data commonly arise because measurement procedures often have detection limits (DL). A DL may represent a floor value, a ceiling value or an interval where precise quantitative levels cannot be determined. For example, exposure assessment for nuclear workers relied on radiation film badges which record radiation levels only above a fixed minimum, due to limits in film photosensitivity (Gilbert et al. 1996; Kerr 1994). Investigators encountered ceiling levels of particle bound polycyclic aromatic hydrocarbons in rural Chinese dwellings when values exceeded 60,000 ng/m^3, the upper limit of the measurement protocol (Ligman et al. 2004).

Although values below or above a DL are "missing", data are not missing at random in the usual sense, since their absence reflects levels of exposure. This type of missing data is called "nonignorable missing", and the simple exclusion of such "interval-measured" data can bias results (Little and Rubin 1987; Schafer 1997).

Analytic procedures for environmental measurement data with DLs are often presented in the context of environmental monitoring where the primary goal is estimation of distributional parameters when numbers of measurements are limited(Gleit 1985; Haas and Scheff 1990; Helsel 1990; Persson and Rootzen 197; Singh and Nocerino 2002; Travis and Land 1990). In epidemiologic studies, measurement data are used to characterize exposures of study subjects, and are typically employed in two ways. Measurement data are used to develop regression models to examine the relationship between a measured value (dependent variable) and covariates (independent variables). Measurement data are also used as covariates in a risk analysis to estimate the relationship between a binary disease outcome and exposure measures and other factors. In the current paper, we focus on the first application, namely, the regression of an exposure measurement on covariate factors. The use of measurements with DLs in risk regression of will be considered in another paper.

Investigators apply various strategies for measurement data with DLs, including replacement of measurements below a DL with a single value, e.g., DL, DL/2 or DL/$\sqrt{2}$ (Helsel 1990; Hornung and Reed 1990). Less frequently, measurements below a DL are assigned a value of zero. However, unless such measurements indicate a true zero exposure, this latter strategy clearly distorts results, and is not considered further. If the distribution of the measurement data is known, e.g., measurements are log-normally distributed, then an alternative strategy replaces values below the DL with expected values of the missing measurements, conditional on being less than the DL (Garland et al. 1993; Gleit

1985). For measurement Z and detection limit DL, we denote this value E[Z | Z<DL]. Calculation of the conditional expected value requires the investigator to either know or estimate parameters of the measurement distribution.

Substitution schemes like those described above are simple, since one value replaces all measurements below the DL, and, except for E[Z | Z<DL], do not consider distributional assumptions. However, since a single value represents all measurements below the DL, parameter estimates and their variances are likely biased, unless the proportion is small, which potentially distorts inference. This limitation led to a single impute "fill-in" method (Helsel 1990; Moschandreas et al. 2001b; Moschandreas et al. 2001a). An investigator first characterizes the form of the distribution and estimates its parameters, then assigns randomly sampled values below the DL from the estimated distribution. Fill-in values along with measured values above the DL are then used in analyses. With appropriate estimation techniques, this approach accommodates multiple DLs.

As described in(Helsel 1990) and applied in(Moschandreas et al. 2001b), the fill-in method did not include complex modeling of regression factors. In addition, while the fill-in approach assigned random values from an appropriate distribution, it did not account for the variability of the imputation process, since the inserted values are not real data. In this paper, we illustrate methods for epidemiologic data that account for measurements with DLs, using data from a case-control study of non-Hodgkin lymphoma (NHL) (Colt et al. 2004). The example evaluates the relationship between concentrations of pesticide analytes in carpet dust and use of pesticide products in and around the home. We restrict analysis to control subjects, with adjustment for study design factors.

Example data from a case-control study of NHL and pesticides

The principal exposure of the general population to pesticides occurs in the home (Nigg et al. 1990) as the result of indoor use, track-in or drift from outdoors, intrusion of vapors from foundation treatments, or take-home contamination from occupational use (Bradman et al. 1997; Lewis et al. 1999; Lewis et al. 2001). Pesticide residues are retained in carpets, migrating into the underlying foam pad, and may persist for months or years.

Data source. We consider data from controls from a multi-center, population-based case-control study of NHL, conducted in the Detroit, Michigan metropolitan area, the state of Iowa, Los Angeles County, California, and the Seattle, Washington metropolitan area (Colt et al. 2004). Controls include 1,057 residents between the ages of 20 and 74, frequency matched to cases on age, gender, race, and study area, with an over-sampling of African American subjects in Los Angeles and Detroit.

Interviewers collected information from respondents on lifetime residential history, and the frequency and form of pesticides used to treat various types of pests (e.g., flying insects, crawling insects, lawn weeds, etc.). Interviewers obtained vacuum cleaner bags from 95 percent of subjects who used their vacuum cleaners within the past year and owned at least half of their carpets or rugs for five years or more. Bags were shipped in insulated boxes by overnight mail to Southwest Research Institute (San Antonio, TX) and placed in freezers. Samples were collected and analyzed for 513 control subjects.

Measurement of carpet dust. The protocol for the collection and measurement of dust samples has been described previously (Colt et al. 2004). Briefly, dust samples were sieved through a 100-mesh sieve to obtain the fine (<150 μm) dust, prior to extraction and analysis. Neutral extractions were carried out for 25 pesticides (18 insecticides, 6 herbicides, and ortho-phenylphenol), 7 polycyclic aromatic hydrocarbons, and 5 polychlorinated biphenyl congeners. Acid extractions were carried out for four herbicides and pentachlorophenol. Extracts were analyzed using gas chromatography/mass spectrometry (GC/MS) in selected ion monitoring mode. Analyte amounts were quantified using the internal standard method. In the full study, GC/MS analysts were blinded to disease status.

After analyzing about half of the samples, investigators began monitoring additional ions for some neutral analytes to clarify identification at low levels, resulting in raised DLs for 14 pesticides. DLs were also raised when less than two grams of dust were available. An additional problem with some dust samples involved the presence of interfering compounds (i.e., compounds that co-eluded with the target analyte), creating uncertainty and prohibiting assignment of specific concentrations.

There were three scenarios for which analysts could provide concentrations only within an interval, which we accommodated by defining a lower bound (LB) and an upper bound (UB) of possible values. If the analyte was not detected and no interferences were present (type I), the LB was set to zero and the UB was set to the specified DL. If there was an interfering compound but insufficient evidence for the presence of the target analyte (type II), the GC/MS analyst reported the result as a nondetect with a detection limit equal to the entire peak of the co-eluting compounds. We set the LB to zero and the UB to 20% of the raised peak, or the DL, which ever was larger. If the target analyte and the interference were both present (type III), the analyst reported an "elevated detect" with a concentration equal to the entire peak of the co-eluting compounds. We set the LB bound to the maximum of 20% of the recorded peak, or the DL, and the UB to the maximum of 90% of the reported peak, or the DL, resulting in an interval bounded away from zero.

For ease of presentation, we allow the replacement of measurements below the DL with DL/2 (which applies to missing data types I and II) to refer more generally to the replacement with (LB+UB)/2 (which applies to missing data types I, II and III).

Methods and analysis

Preliminary analysis indicates that measurement data are consistent with a log-normal distribution. If Z denotes the measured value of an analyte and is log-normally distributed, denoted $Z \sim LN(\mu, \sigma^2)$, then by definition log(Z) is a normal random variable with mean μ and variance σ^2, denoted $\log(Z) \sim N(\mu, \sigma^2)$ (Singh et al. 1997). Suppose $X=(X_o,\ldots,X_K)^t$ is a column vector of covariates, with $X_o=1$, and

$\beta=(\beta_o,\ldots,\beta_K)^t$ is a column vector of regression parameters, where "t" denotes vector transpose. If data are complete, then a linear regression equation has the form $\log(Z) = \beta^t X + \epsilon$, where $\epsilon \sim N(0,\sigma^2)$. For each X, the model implies that Z is log-normally distributed with mean $\beta^t X$, i.e., $Z \sim LN(\beta^t X, \sigma^2)$.

Regression analysis in control data. We evaluate the association between analyte concentration and pesticide use by fitting a linear regression model of the logarithm of the analyte level on subject characteristics. Regression (independent) covariates include indicator variables for season of sample collection, presence of oriental rugs, study center, sex, age (<45, 45-64, ≥65 years), race (African American, Caucasian, other), type of home (single family, townhouse/duplex/apartment, other), year of home construction (<1940, 1940-1959, 1960-1979, ≥1980), and educational level (<12, 12-15, ≥16 years). As in Colt et al. (Colt et al. 2004), covariates vary slightly with analyte. Models also include five variables describing the use of insect treatment products: ever/never used products to treat for crawling insects, flying insects, fleas/ticks, termites, and lawn/garden insects. We use data from current homes only.

Regression analysis is hampered by the presence of measurements known only within bounds. We assume that the probability distributions of measurements below the DL (more precisely within the LB and UB interval) depend only on observed data, i.e., the interval-measured concentrations arise from the same distributions that generate the measured values. Let $F(\cdot)$ be the cumulative distribution function and $f(\cdot)$ the probability density function for a log-normal random variable. Suppose $X_i = (X_{i\,o},\ldots,X_{i\,K})^t$ is the covariate vector for the i^{th} of $i=1,\ldots,n$ subjects. LB_i and UB_i are recorded for $i=1,\ldots,n_o$ individuals, while a specific Z_i measurement is recorded for $i=n_o+1,\ldots,n_o+n_1$ individuals. LB and UB are subscripted to allow different DLs. Using a Tobit regression approach (Gilbert 1987; Persson and Rootzen 197; Tobin 1958), the log-likelihood function has the form

$$L(\beta,\sigma^2) = \sum_{i=1}^{n_o} \log[F(Z_{UB_i};\beta^tX_i,\sigma^2) - F(Z_{LB_i};\beta^tX_i,\sigma^2)] + \sum_{i=n_o+1}^{n_o+n_1} \log[f(Z_i;\beta^tX_i,\sigma^2)] \quad (1)$$

The first summand derives from the n_o interval-measured values and involves the difference of the cumulative distribution function F evaluated at UB and at LB, i.e., the probability the measurement lies between the LB and UB. The second summand derives from the n_1 detected values. Maximum likelihood estimates (MLE) for β and their covariance matrix are obtained by maximizing equation (1) and computing the inverse information matrix using standard methods.

Imputation of missing concentrations. If the goal is the evaluation of pesticide use and analyte levels in carpet dust, represented by the β parameters, then the Tobit regression of equation (1) is sufficient and no imputation is required. If there is the need for further analysis or for graphical display, it is useful to generate values for measurements below DLs. We consider several different approaches, including inserting DL/2, inserting $E[Z \mid Z<DL]$, or using a single or multiple imputation (Little and Rubin 1987).

A multiple imputation procedure is carried out as follows. Using all data (measured concentrations, missing data types I-III, and covariates), we create the log-likelihood function (1), solve for the MLEs of β and σ^2, denoted $\hat{\beta}$ and $\hat{\sigma}^2$, and impute a value by randomly sampling from a log-normal distribution with the estimated parameters. However, in selecting fill-in values we cannot ignore that $\hat{\beta}$ and $\hat{\sigma}^2$ are themselves estimates with uncertainties. We therefore do not use $\hat{\beta}$ and $\hat{\sigma}^2$ for the imputation, but rather $\tilde{\beta}$ and $\tilde{\sigma}^2$, which are estimated from a bootstrap sample of the data (Efron 1979). Bootstrap data are generated as described below by sampling with replacement, and represent a sample from the same universe as the original data. We repeat the process to create multiple data sets, which are then independently analyzed and combined in a way that accounts for the imputation. Differences in regression results in the multiple data sets reflect variability due to the imputation process.

This procedure however omits a source of variability. We have tacitly assumed that the LB and UB are fixed and known in advance. When there are no interfering compounds (missing type I), the assumption is justified since the DL is determined prior to the GC/MS dust analysis. When there are interfering compounds (missing types II and III), the assumption cannot be fully justified since the bounds depend on the amount of interference and therefore are random. In the NHL data, we assume this uncertainty is small relative to other uncertainties.

The imputation proceeds as follows.

Step 1: Create a bootstrap sample and obtain estimates $\tilde{\beta}$ and $\tilde{\sigma}^2$ based on equation (2). Bootstrap data are generated by sampling with replacement n times from the n subjects. Sampling "with replacement" selects one record at random, then "puts it back" and selects a second record. After n repetitions, some subjects are selected multiple times, while other subjects are not selected at all. If w_i is the number of times the i^{th} subject is sampled, then the log-likelihood function for the bootstrap data is

$$L(\mu,\sigma^2) = \sum_{i=1}^{n_0} w_i \log[F(z_{UB_i};\beta^t X_i,\sigma^2) - F(z_{LB_i};\beta^t X_i,\sigma^2)] + \sum_{i=n_0+1}^{n_0+n_1} w_i \log[f(z_i;\beta^t X_i,\sigma^2)] \quad (2)$$

Step 2: Impute analyte values based on sampling from $LN(\tilde{\beta}^t X, \tilde{\sigma}^2)$. For the i^{th} subject, assign the value:

$$F^{-1}\{\mathrm{Unif}[F(LB_i;\tilde{\mu},\tilde{\sigma}^2)], F(UB_i;\tilde{\mu},\tilde{\sigma}^2)]; \tilde{\mu},\tilde{\sigma}^2\} \qquad (3)$$

This quantity consists of various elements. $F(LB_i; \tilde{\beta}^t X, \tilde{\sigma}^2)$ and $F(UB_i; \tilde{\beta}^t X, \tilde{\sigma}^2)$ are the cumulative probabilities at UL_i and UB_i, respectively, based on parameters $\tilde{\beta}$, $\tilde{\sigma}^2$. Both values lie between zero and one. Select randomly from a uniform distribution on the interval [a,b], denoted Unif[a,b], in particular, the interval $[F(LB_i; \tilde{\beta}^t X_i, \tilde{\sigma}^2), F(UB_i; \tilde{\beta}^t X_i, \tilde{\sigma}^2)]$. The inverse cumulative distribution function, $F^{-1}(\cdot)$ is the required imputed value in original units between LB_i and UB_i. Repeat using the same $\tilde{\beta}$, $\tilde{\sigma}^2$ for each missing value. Detected values are not altered.

Step 3: Repeat step 1 and step 2 to create M plausible (or "fill-in") data sets. Remarkably, M need not be large, and a recommended value is between 3 and 5, with larger values if greater proportions of data are missing (Little and Rubin 1987; Rubin 1987). We select M=10 to fully account for the variance from the imputation.

Step 4: Fit a regression model to each of the M data sets and obtain M sets of parameter estimates and covariance matrices. Combine the M sets of estimates to account for the imputation (Little and Rubin 1987; Schafer 1997). The imputation procedure results in confidence intervals, which are wider than the single-imputation, fill-in approach.

Simulation study. We conduct a simulation study, using a simple regression model with zero intercept and no covariates, to evaluate the imputation approaches and the effects of the proportion of data below the DL and sample size. We generate data sets of size n by sampling from a log-normal distribution with parameters (μ,σ^2), and defined the DL such that in expectation p percent of the samples falls below the DL, i.e., DL = $F^{-1}(p;\mu,\sigma^2)$. The simulation involves 5,000 independent data sets for each set of parameters. We compared five approaches.

(i) Direct estimation (Tobit regression) of MLEs $(\hat{\mu},\hat{\sigma}^2)$ using equation (1).

(ii) Multiple imputation with allowance for uncertainty in model parameters.

(iii) Single imputation based on a random fill-in value for each datum below the DL, using MLEs $(\hat{\mu},\hat{\sigma}^2)$ from equation (1).

(iv) Insertion of DL/2 for all data below the DL.

(v) Insertion of E[Z | Z<DL] for data below the DL with the expected value based on the MLEs $(\hat{\mu},\hat{\sigma}^2)$ from equation (1).

For approaches (ii)-(v), estimators are the mean and variance of the logarithm of the observed and imputed data, with adjustment for multiple imputation in (ii). We compare results to estimates based on complete data.

For the NHL example, we use SAS to generate bootstrap samples, fit linear regressions (PROC REG), solve log-likelihood equations (1) and (2) (PROC LIFEREG), and combine results from multiple data sets (PROC MIANALYZE) (SAS Institute Inc. 2001. The SAS System for Windows, Version 8.2. Cary, NC, USA.) The simulation was conducted using MATLAB (The MathWorks, Inc. 2004. MATLAB, The Language of Technical Computing, Version 7.0. Natick, MA 01760-1500.)

Results

We limit results to four insecticides, propoxur and carbaryl, both carbamate insecticides, chlorpyrifos, an organophosphate, and a-chlordane, an organochlorine, which exhibited various types and proportions of missing data.

Regression analysis in control subjects. After omitting subjects missing questionnaire data, there are 478 control subjects with carpet dust measurements and all regression variables. The percentages of measurements below DLs or known only within bounds vary from 25.7 percent for propoxur to 67.0 percent for carbaryl (Table 1). The arithmetic mean (AM), geometric mean (GM) and geometric standard deviation (GSD), with fill-in imputations for interval-measured values, indicate that concentrations for the individual analytes varied substantially. For carbaryl and a-chlordane, the GM falls within the range of missing data. Figure 1, panels A and B, show quantile plots for measurements of propoxur and carbaryl, and reveal good concordance with a log-normal distribution. Panels A and B show values used for imputation based on DL/2, denoted by stars, and the conditional expected value, denoted by triangles. For carbaryl, DL/2 values are nearly twice the conditional expected values. By construction, the fill-in values conform to the estimated distribution.

Table 2 shows proportional effects of the use of the insecticide products in and around the home for direct estimation of regression parameters (Tobit regression), the multiple imputation approach, the replacement of missing concentrations by DL/2 and $E[Z \mid LB<Z<UB]$, and a single set of fill-in values. Results differ slightly from those reported in Colt (Colt et al. 2004) due to differences in regressor variables. For the fill-in approach, we impute missing values using a model with regression variables (denoted "yes") and without regression variables except for an intercept variable (denoted "no").

In several instances, estimates for the various products differ substantially, particularly for analytes with a high percentage of missing data. The multiplicative standard errors for the single imputation approaches (i.e., inserting DL/2, $E[Z \mid LB<Z<UB]$, or a fill-in value) are smaller than standard errors from the multiple imputation approach and direct estimation. The smaller standard errors result from an inadequate account of missing data and result in confidence intervals (CI) which are too narrow and inflated Type I error rates. Table 2 shows several variables that do not achieve traditional significance levels when imputation is taken into account. In some instances, there are marked differences in estimates. Estimated increases in carpet dust levels of a-chlordane among subjects treating for termites are 2.6 and 3.1-fold based on DL/2 insertion and fill-in methods, respectively, and 3.7-fold based on multiple imputation and direct estimation approaches.

Comparing the two fill-in approaches, standard errors are smaller when the covariate information is included, than when covariate information is omitted.

Fill-in values are obtained from regression models by sampling from $LN(\hat{\beta}^t X, \hat{\sigma}^2)$. Panels C and D in Figure 1 show quantile plots of residuals, i.e., $\exp[\log(Z) - \hat{\beta}^t X]$ for each subject. While GMs of the residuals are close to the expected value of 1.0 for the error distributions, plots suggest a slight under-prediction at extreme values for propoxur and carbaryl.

Simulation study. For the simulation study, we set $\mu = 0$ and $\sigma^2 = 1$ without loss of generality, and present results for n = 50, 100, 200, and 400 and with DLs such that the expected proportions of values below the DL are p=10, 30, 50, and 70 percent. With 5,000 repetitions, the standard error for coverage of 95 percent CIs is 0.003. Table 3 shows that estimates of μ based on Tobit regression, multiple imputation, and single impute fill-in approaches are generally unbiased. Insertion of DL/2 or E[Z | Z<DL] results in substantial bias unless the proportion of missing data is small, 10 percent or less. The table also shows coverage of the 95 percent CI for the estimate of μ. In comparison with complete data, Tobit regression and the multiple imputation approaches are the only methods which achieve nominal coverage over a broad range of simulation parameters, although the multiple imputation begins to degrade when more than about 50 percent of the measurements are below DLs. The single imputation approach results in anomalous CIs, when about 30 percent or more of the data are below DLs.

Discussion

Results of our analysis of use of pesticide products in and around the home and pesticide residues in carpet dust, and of the simulation study suggest that the method of imputation of missing environmental measurement data can substantially impact estimation of effects and statistical inference. The practice of inserting a single value, such as DL/2 or the conditional expected value E[Z | Z<DL] or by analogy DL/√2, is ill-advised unless there are relatively few measurements below detection limits. The use of a single imputation to fill-in missing data is unbiased or minimally biased quite generally, but suffers from inaccurate estimates

of variance and, as a consequence, CIs that are too narrow, particularly when missing data exceed about 30 percent. The best protection against biased inference in the presence of nonignorable missing data is the use of multiple imputation, although with a high proportion of values below the DL a large number of measurements are needed. It is worth reiterating however that multiple imputation is necessary only if explicit values are needed for measurements below DLs. If the purpose is estimation of regression parameters, then procedures for truncated data, such as Tobit regression, are nominal (Little and Rubin 1987).

In environmental monitoring, estimation of distributional parameters is often problematic due to limited numbers of measurements and an inability to evaluate distributional forms precisely. With 5-15 measurements, MLEs can be biased (Gleit 1985), suggesting the need for more robust approaches (Helsel 1990). With epidemiologic data, which usually include hundreds or thousands of measurements, MLEs are unbiased and fully efficient (Gilliom and Helsel 1986), and more detailed regression analyses are feasible.

When analyzing environmental data on pesticides, Moschandreas et al used a fill-in imputation approach that applied the "best fitting" probability distribution for values above a detection limit (Helsel 1990; Moschandreas et al. 2001b; Moschandreas et al. 2001a), although Helsel and Hirsch (Helsel and Hirsch 2004) had cautioned that the approach should be used primarily for estimating summary statistics. The approach we outline permits multiple DLs, incorporates regression parameters, and applies multiple imputation to account correctly for interval measured data and to allow unbiased inference. However, our simulation study suggests that the fill-in approach may be quite adequate when measurements below the DL account for less than about 30 percent of the data.

The Tobit regression and multiple imputation approaches assume that the limits of detection are fixed and known in advance. In our example, we are justified in assuming DLs are fixed for type I missing measurements, but not for types II and III missing data where limits of detection depend on the amount of interfering compounds and are random variables. If the DL is not known, an estimate of its value is the minimum order statistic of the observed measurements, that is, the smallest measured value. Simulations suggest that for a random DL estimates remain unbiased, but variances are underestimated (Zuehlke 2003). Thus, CIs in Table 2 may be too narrow. However, relative to other sources of uncertainty that arise in the collection and handling of carpet dust samples, and the accuracy of questionnaire information, additional uncertainty induced by random DLs for type II and III missing values is likely small.

Environmental data are frequently well approximated by a log-normal distribution, and our data on concentrations of pesticide analyte in carpet dust are consistent with this assumption. Equations (1) and (2) remain valid for more general distributions, although estimation of parameters may be more problematic and necessitate potentially computer-intensive search algorithms. Validity of parameter estimates and their variances depend of course on the correct choice of error distribution. Our simulation study was based on a correct distributional form;

however, misspecification of the probability model can lead to markedly biased results (Paarsch 1984). In the absence of knowledge about the error distribution, semiparametric and nonparametric methods have been proposed (Austin 2002a; Chay and Powell 2001; DiNardo and Tobias 2001). Bayesian approaches have also been suggested in the Tobit regression context (Austin 2002b). A Reviewer suggested considering the set of measurements of a subject as a vector of multivariate outcomes, so that the covariance structure among the analytes could provide information for the imputation process. In our example, this requires the assumption that the logarithms of the measurements are multivariate normally distributed. The suggestion however adds complexity as the number of analytes increases, and additional work is needed to evaluate its practical feasibility.

The motivation for this work arose from the analysis of pesticide analytes in carpet dusts and use of pesticide products in and around the home. However, data with DLs arise in a variety of settings including, upper DLs from healthcare related questionnaire data (Austin 2002a) and psychological profile scores, such as the Fagerstrom Test for Nicotine Dependence (Fagerstrom and Schneider 1989; Heatherton et al. 1991), and lower DLs in radiation film badge measurements (Gilbert et al. 1996; Kerr 1994).

In summary, with epidemiologic data, our analyses indicate that unless there are very few measurements below DLs (less than 5-10 percent), inserting DL/2, $E[Z \mid Z<DL]$, or any single value to impute missing measurement data is not advisable. Further, inserting a randomly selected fill-in value is also inadvisable,

unless the proportion of missing data is less than about 30 percent. Multiple imputation of missing data is the best approach of ensuring unbiased estimates of effects and nominal confidence intervals.

References

Austin PC. 2002a. A comparison of methods for analyzing health-related quality-of-life measures. Value Health 5:329-337.

Austin PC. 2002b. Bayesian extensions of the Tobit model for analyzing measures of health status. Med Decis Making 22:152-162.

Bradman MA, Harnly ME, Draper W, Seidel S, Teran S, Wakeham D, et al. 1997. Pesticide exposures to children from California's Central Valley: Results of a pilot study. J Expo Anal Environ Epidemiol 7:217-234.

Chay KY, Powell JL. 2001. Semiparametric censored regression models. J Econ Perspect 15:29-42.

Colt JS, Lubin J, Camann D, Davis S, Cerhan J, Severson RK, et al. 2004. Comparison of pesticide levels in carpet dust and self-reported pest treatment practices in four US sites. J Expo Anal Environ Epidemiol 14:74-83.

Desrosiers M, Schauer DA. 2001. Electron paramagnetic resonance (EPR) biodosimetry. Nucl Instrum Meth B 184:219-228.

DiNardo J, Tobias J. 2001. Nonparametric density and regression estimation. J Econ Perspect 15:11-28.

Efron B. 1979. Bootstrap methods; another look at the jack-knife. Ann Statist 7:1-26.

Fagerstrom KO, Schneider NG. 1989. Measuring nicotine dependence - a review of the Fagerstrom tolerance questionnaire. J Behav Med 12:159-182.

Garland M, Morris JS, Rosner BA, Stampfer MJ, Spate VL, Baskett CJ, et al. 1993. Toenail trace-element levels as biomarkers - reproducibility over a 6-year period. Cancer Epidemiol Biomarkers Prev 2:493-497.

Gilbert ES, Fix JJ, Baumgartner WV. 1996. An approach to evaluating bias and uncertainty in estimates of external dose obtained from personal dosimeters. Health Phys 70:336-345.

Gilbert RO. 1987. Statistical Methods for Environmental Pollution Monitoring. New York:Van Nostrand Reinhold.

Gilliom RJ, Helsel DR. 1986. Estimation of distributional parameters for censored trace level water quality data, 1. Estimation techniques. Water Resour Res 22:135-146.

Gleit A. 1985. Estimation for small normal data sets with detection limits. Environ Sci Technol 19:1201-1206.

Haas CN, Scheff PA. 1990. Estimation of averages in truncated samples. Environ Sci Technol 24:912-919.

Hatch M, Thomas D. 1993. Measurement issues in environmental epidemiology. Environ Health Perspect 101:49-57.

Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO. 1991. The Fagerstrom test for nicotine dependence - a revision of the Fagerstrom tolerance questionnaire. Br J Addict 86:1119-1127.

Helsel DR. 1990. Less than obvious - statistical treatment of data below the detection limit. Environ Sci Technol 24:1766-1774.

Helsel DR, Hirsch RM. 1991. Techniques of Water-Resources Investigations of the United States Geological Survey. Book 4, Hydrologic Analysis and Interpretation. Chapter A3. Statistical Methods in Water Resources. Available: http://water.usgs.gov/pubs/twri/twri4a3/ [accessed 13 August 2004].

Hornung RW, Reed LD. 1990. Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg 5:46-51.

Kerr GD. 1994. Missing dose from mortality studies of radiation effects among workers at Oak-Ridge-National-Laboratory. Health Phys 66:206-208.

Lewis RG, Fortune CR, Blanchard FT, Camann DE. 2001. Movement and deposition of two organophosphorus pesticides within a residence after interior and exterior applications. J Air Waste Manage Assoc 51:339-351.

Lewis RG, Fortune CR, Willis RD, Camann DE, Antley JT. 1999. Distribution of pesticides and polycyclic aromatic hydrocarbons in house dust as a function of particle size. Environ Health Perspect 107:721-726.

Ligman B, Shaughnessy R, Kleinerman R, Lubin J, Fisher E, Wang ZY, et al. Indoor air pollution characterization of underground dwellings in China., 1997, Blacksburg:VPI and State University Press,2004;51-56.

Little RJA, Rubin DB.1987. Statistical Analysis with Missing Data. New York:John Wiley & Sons, Inc.

Moschandreas DJ, Karuchit S, Kim Y, Ari H, Lebowitz MD, O'Rourke MK, et al. 2001a. On predicting multi-route and multimedia residential exposure to chlorpyrifos and diazinon. J Expo Anal Environ Epidemiol 11:56-65.

Moschandreas DJ, Kim Y, Karuchit S, Ari H, Lebowitz MD, O'Rourke MK, et al. 2001b. In-residence, multiple route exposures to chlorpyrifos and diazinon estimated by indirect method models. Atmos Environ 35:2201-2213.

Nigg N, Beier RC, Carter O, Chaisson C, Franklin C, Lavy T, Lewis RG, Lombardo P, McCarthy JF, Maddy KT. 1990. Exposure to pesticides. In: Vol XVIII: The Effects of Pesticides on Human Health (Baker SR, Wilkinson CF, eds) Princeton, NJ:Princeton Scientific, 35-130.

Paarsch HJ. 1984. A monte-carlo comparison of estimators for censored regression-models. J Econ 24:197-213.

Persson T, Rootzen H. 1977. Simple and highly efficient estimators for a type I censored normal sample. Biometrika 64:123-128.

Rubin DB. 1987. Multiple Imputation for Nonresponse in Surveys. New York:J. Wiley & Sons.

Schafer JL. 1997. Analysis of Incomplete Multivariate Data. New York:Chapman & Hall/CRC.

Sim M. 2002. Case studies in the use of toxicological measures in epidemiological studies. Toxicology 181:405-409.

Singh A, Nocerino J. 2002. Robust estimation of mean and variance using environmental data sets with below detection limit observations. Chemometr Intell Lab 60:69-86.

Singh AK, Singh A, Engelhardt M. 1997. The Lognormal Distribution in Environmental Applications. Washington, D.C:U.S.Environmental Protection Agency, Office of Solid Waste and Emergency Response.

Tobin J. 1958. Estimation of relationships for limited dependent variables. Econometrica 26:24-36.

Travis CC, Land ML. 1990. Estimating the mean of data sets with nondetectable values. Environ Sci Technol 24:961-962.

Woodward A, Al Delaimy W. 1999. Measures of exposure to environmental tobacco smoke - Validity, precision, and relevance. Uncertainty Risk Assessment Environ Occup Hazards 895:156-172.

Zuehlke TW. 2003. Estimation of a Tobit model with unknown censoring threshold. Appl Econ 35:1163-1169.

Table 1: Percentage of measurements below detection limits or known only within bounds and arithmetic means (AM), geometric means (GM) and geometric standard deviations (GSD), based on fill-in values from a single imputation. Data on 478 control subjects.

	Measurements known only within bounds [a]								
	Type I		Type II		Type III		ng/g dust		
Insecticide	%	Range	%	Range	%	Range	AM	GM	GSD
Propoxur	21.1	0-46.0	2.9	0-65.0	1.7	21.1-75.7	456.6	65.6	6.0
Carbaryl	37.9	0-260.0	11.1	0-268	18.0	20.7-694.8	1503.0	64.0	14.0
Chlorpyrifos	28.2	0-77.4	0.2	0-20.9	0.0	–	893.1	105.6	8.3
a-Chlordane	60.9	0-44.7	0.0	–	0.4	20.8-29.1	90.7	11.6	8.0

[a] Types of missing measurements are: no analyte detected and no interfering compound (I), no analyte detected but with an interfering compound present (II), and analyte and interfering compounds both present. The range for the detection limits reflects the minimum of lower bounds and the maximum of upper bounds for the non-detected measurements.

Table 2: Proportional increase in analyte concentration in carpet dust in nano-grams per gram dust (ng/g) for selected uses [a].

	Imputation approach [b]		Crawling insects		Flying insects		Fleas/ticks		Termites		Lawn/garden insects	
Insecticide	Method	Adjustment	exp(β)	SE	exp(β)	SE	exp(β)	SE	exp(β)	SE	exp(β)SE	
Propoxur	DL/2	no	1.426[c]	1.167	0.987	1.144	1.231	1.153	1.145	1.219	0.756[c]	1.151
	E[Z | LB<Z<UB]	no	1.432[c]	1.170	0.986	1.147	1.231	1.156	1.135	1.223	0.751[c]	1.154
	Fill-in	no	1.459[c]	1.189	0.966	1.163	1.225	1.173	1.072	1.249	0.737[d]	1.171
	Fill-in	yes	1.511[c]	1.182	1.030	1.157	1.251	1.166	1.209	1.239	0.687[c]	1.165
	Multiple impute	yes	1.487[c]	1.196	1.016	1.165	1.247	1.170	1.082	1.244	0.704[c]	1.173
	Direct estimate	yes	1.503[d]	1.276	0.994	1.235	1.245	1.250	1.090	1.363	0.714	1.249
Carbaryl	DL/2	no	0.853	1.201	0.661[c]	1.173	1.560[c]	1.185	1.129	1.266	1.660[c]	1.183
	E[Z | LB<Z<UB]	no	0.849	1.226	0.629[c]	1.194	1.703[c]	1.208	1.199	1.300	1.746[c]	1.205
	Fill-in	no	0.830	1.311	0.591[c]	1.265	1.812[c]	1.285	1.486	1.417	1.735[c]	1.282
	Fill-in	yes	0.940	1.274	0.432[c]	1.235	2.337[c]	1.252	1.538	1.366	1.779[c]	1.249
	Multiple impute	yes	0.826	1.338	0.508[c]	1.272	2.047[c]	1.313	1.326	1.490	1.950[c]	1.351
	Direct estimate	yes	0.785	1.499	0.512[d]	1.413	2.180[c]	1.452	1.281	1.651	2.115[c]	1.444
Chlorpyrifos	DL/2	no	1.578[c]	1.209	0.779	1.181	1.264	1.182	1.581[d]	1.276	0.759	1.188
	E[Z | LB<Z<UB]	no	1.620[c]	1.218	0.771	1.188	1.300	1.190	1.613[d]	1.288	0.746	1.196
	Fill-in	no	1.917[c]	1.243	0.757	1.210	1.389[d]	1.212	1.669[d]	1.322	0.713[d]	1.219
	Fill-in	yes	1.745[c]	1.244	0.740	1.210	1.383[d]	1.212	1.631[d]	1.323	0.731	1.219
	Multiple impute	yes	1.770[c]	1.252	0.763	1.223	1.401[d]	1.223	1.689[d]	1.336	0.708	1.234
	Direct estimate	yes	1.796[d]	1.378	0.740	1.323	1.392	1.327	1.698	1.492	0.702	1.338
a-Chlordane	DL/2	no	0.966	1.129	0.938	1.112	0.910	1.118	2.626[c]	1.168	1.091	1.117
	E[Z | LB<Z<UB]	no	0.954	1.153	0.925	1.132	0.894	1.140	3.031[c]	1.199	1.110	1.138
	Fill-in	no	1.060	1.230	0.828	1.198	0.868	1.210	3.110[c]	1.303	1.079	1.208
	Fill-in	yes	0.762	1.206	0.927	1.177	0.908	1.188	3.898[c]	1.271	1.293	1.186
	Multiple impute	yes	0.852	1.363	0.915	1.235	0.804	1.202	3.686[c]	1.290	1.169	1.270
	Direct estimate	yes	0.858	1.379	0.919	1.316	0.803	1.339	3.666[c]	1.442	1.211	1.334

[a] Entries are exponentials of parameter estimates (β) and their standard errors (SE) from linear regression models of the logarithm of pesticide analyte on age, sex, race, education, study site, season, and pesticide use variables. Regression models also included year house was built (propoxur, carbaryl, a-chlordane), type of home (carbaryl) and presence of oriental rugs (a-chlordane).

[b] See text for a description of methods. Adjusted imputation accounts for regression variables.

[c] 95 percent confidence interval excludes one.

[d] 90 percent confidence interval excludes one.

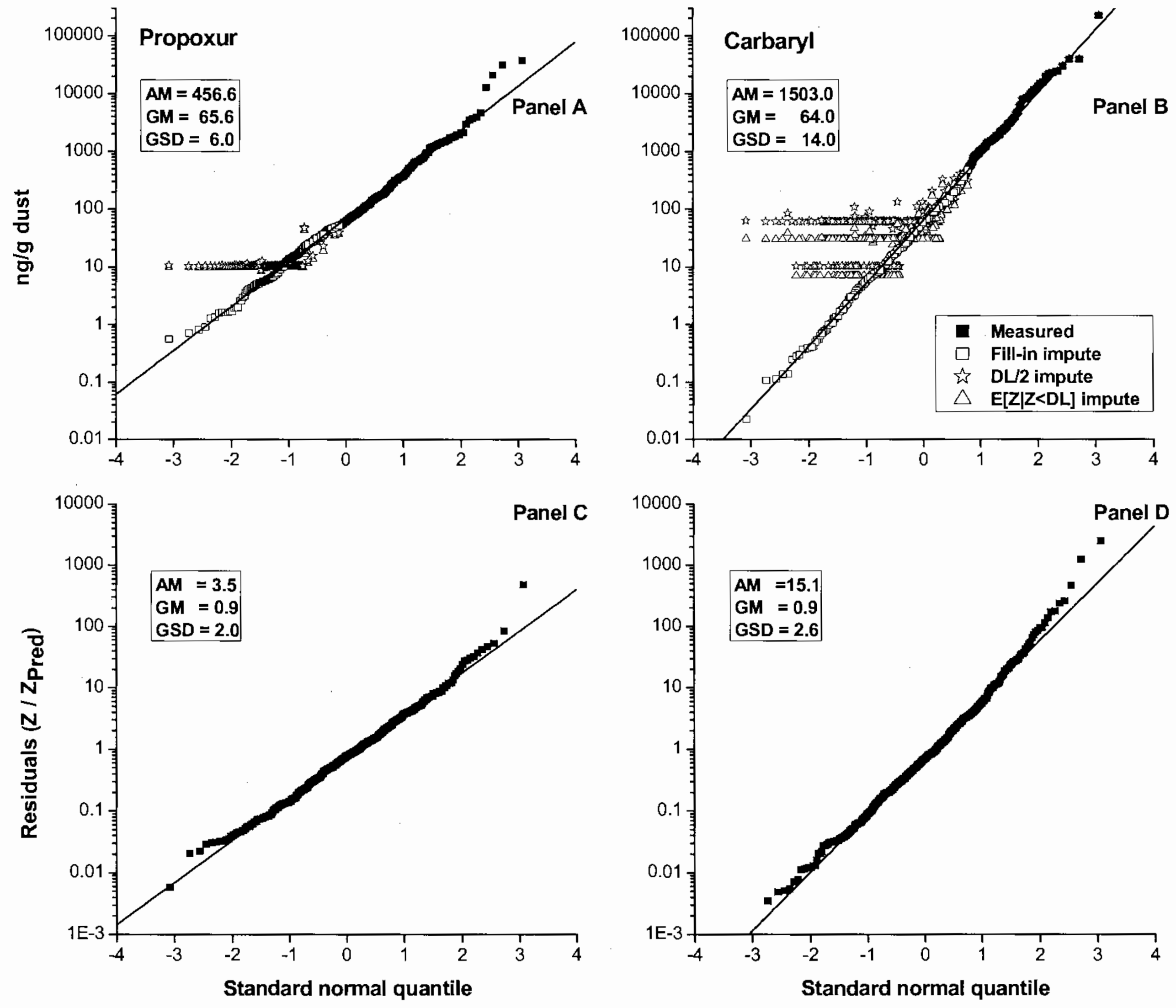
Table 3: Results of simulation study of imputation approaches[a] for log-normally distributed data with μ=0 and σ^2=1 with a detection limit (DL). Entries are means of 5,000 repetitions.

Sample size	% <DL	Complete data	Tobit analysis	Multi impute using $(\tilde{\mu},\tilde{\sigma}^2)$	Single impute using $(\hat{\mu},\hat{\sigma}^2)$	Insert DL/2	Insert E[Z|Z<DL]
				Estimate of μ			
n = 50	10.0	0.002	0.000	-0.003	-0.003	-0.020	0.007
	30.0	0.002	-0.003	-0.003	-0.004	-0.017	0.032
	50.0	0.002	-0.004	-0.003	-0.003	0.052	0.073
	70.0	0.002	-0.006	-0.005	-0.002	0.229	0.143
				Coverage of 95% CI			
	10.0	0.947	0.944	0.943	0.943	0.943	0.942
	30.0	0.947	0.949	0.938	0.928	0.942	0.928
	50.0	0.947	0.953	0.928	0.876	0.938	0.832
	70.0	0.947	0.931	0.895	0.707	0.280	0.520
				Estimate of μ			
n = 100	10.0	0.003	0.002	0.000	0.000	-0.019	0.009
	30.0	0.003	0.001	0.000	0.000	-0.015	0.034
	50.0	0.003	0.000	0.000	-0.001	0.055	0.076
	70.0	0.003	-0.006	-0.004	-0.002	0.232	0.142
				Coverage of 95% CI			
	10.0	0.944	0.945	0.940	0.940	0.943	0.942
	30.0	0.944	0.949	0.938	0.929	0.942	0.914
	50.0	0.944	0.948	0.922	0.870	0.910	0.781
	70.0	0.944	0.940	0.904	0.721	0.036	0.440
				Estimate of μ			
n = 200	10.0	-0.001	-0.002	-0.002	-0.002	-0.023	0.006
	30.0	-0.001	-0.003	-0.003	-0.003	-0.019	0.031
	50.0	-0.001	-0.002	-0.002	-0.002	0.052	0.074
	70.0	-0.001	-0.003	-0.001	-0.002	0.229	0.142
				Coverage of 95% CI			
	10.0	0.952	0.950	0.951	0.950	0.941	0.946
	30.0	0.952	0.955	0.936	0.926	0.940	0.904
	50.0	0.952	0.948	0.925	0.874	0.877	0.708
	70.0	0.952	0.947	0.914	0.725	0.000	0.306
				Estimate of μ			
n = 400	10.0	0.001	0.001	0.001	0.001	-0.021	0.008
	30.0	0.001	0.000	0.000	0.000	-0.017	0.034
	50.0	0.001	0.001	0.001	0.001	0.053	0.076
	70.0	0.001	0.000	0.000	0.000	0.230	0.144
				Coverage of 95% CI			
	10.0	0.954	0.954	0.952	0.951	0.931	0.949
	30.0	0.954	0.948	0.938	0.928	0.941	0.874
	50.0	0.954	0.954	0.927	0.880	0.776	0.545
	70.0	0.954	0.947	0.914	0.723	0.000	0.128

[a] Parameter estimation using observed data with DLs (Tobit analysis), $(\hat{\mu},\hat{\sigma}^2)$, multiple imputation with allowance for uncertainty in model parameters $(\tilde{\mu},\tilde{\sigma}^2)$, a single imputation using $(\hat{\mu},\hat{\sigma}^2)$, the insertion of DL/2, and insertion of the expected value conditional on being below the DL, E[Z|Z<DL].

Figure legend:

Figure 1: Plots under a log-normal distribution of quantiles of environmental measurements of propoxur and carbaryl (panels A and B), and of regression residuals of measurements (Z) and predicted values (Z_{Pred}) after accounting for covariates (panels C and D). The arithmetic means (AM), geometric means (GM), and geometric standard deviations (GSD) are computed from imputed data.


Propoxur
AM = 456.6
GM = 65.6
GSD = 6.0
Panel A
ng/g dust
100000
10000
1000
100
10
1
0.1
0.01
-4
-3
-2
-1
0
1
2
3
4
Carbaryl
AM = 1503.0
GM = 64.0
GSD = 14.0
Panel B
■ Measured
□ Fill-in impute
☆ DL/2 impute
△ E[Z|Z<DL] impute
Panel C
AM = 3.5
GM = 0.9
GSD = 2.0
Residuals (Z / Z_Pred)
10000
1E-3
Standard normal quantile
Panel D
AM =15.1
GM = 0.9
GSD = 2.6
Standard normal quantile